

# Integrating MQ with NUMA

Hubertus Franke, Shailabh Nagar, Mike Kravetz

IBM

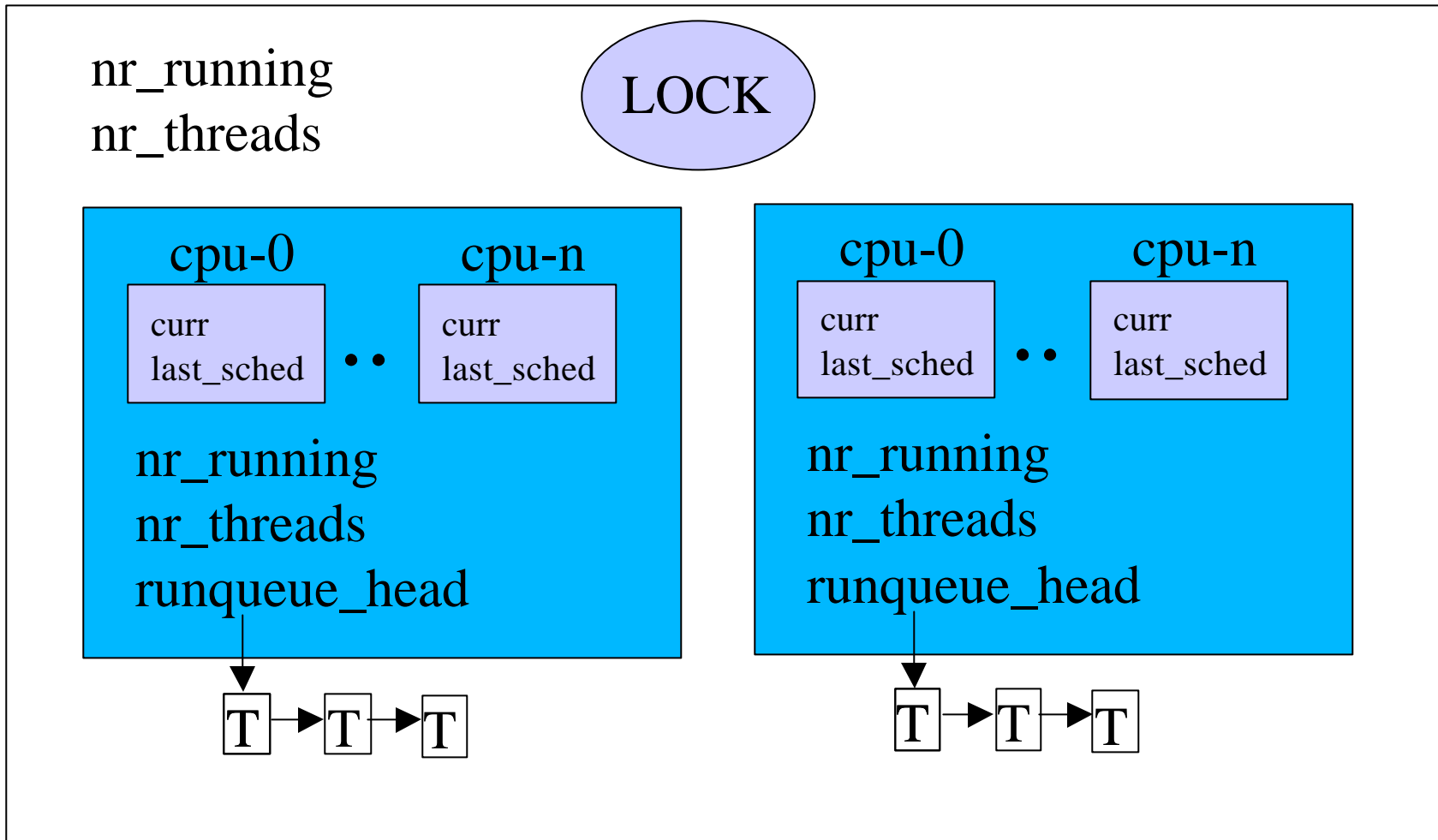
Andrea Arcangeli

Suse Inc.

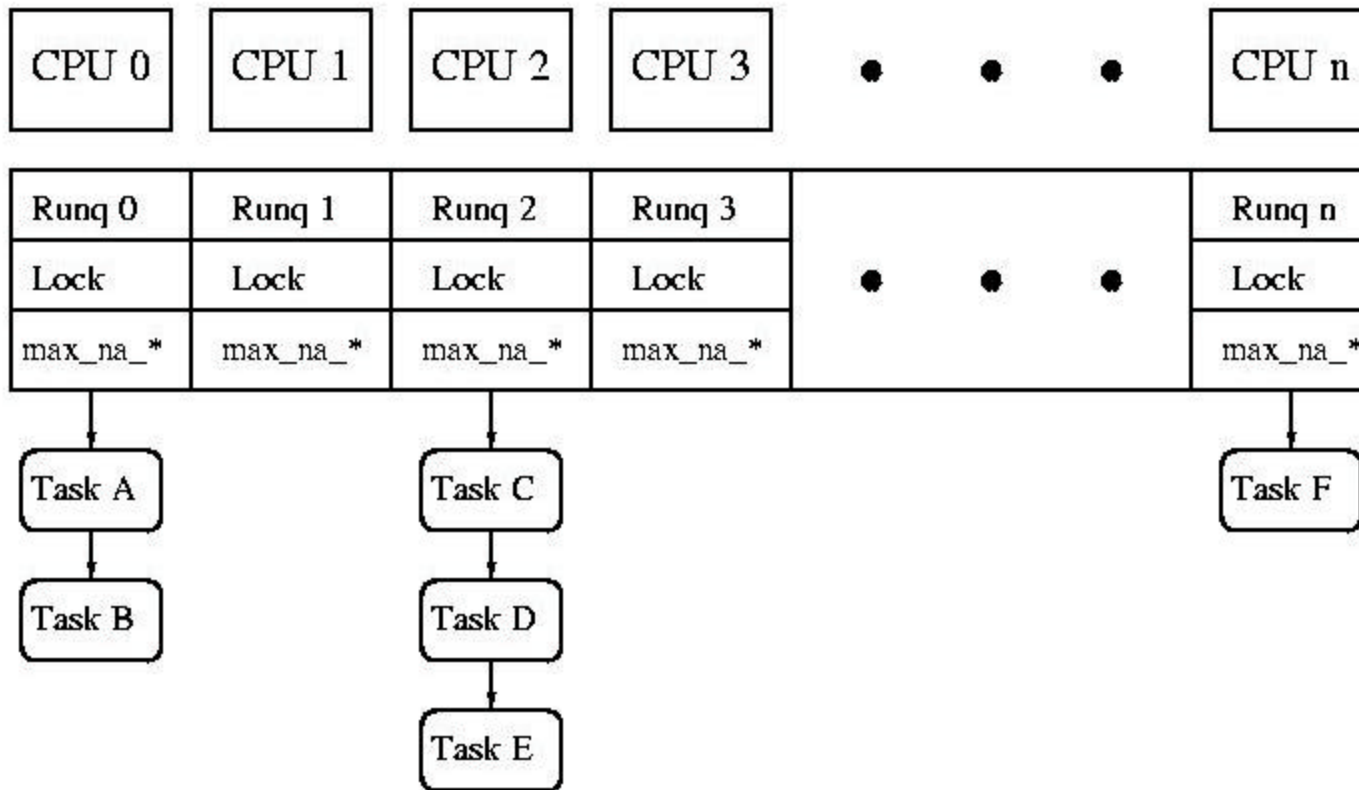
# Overview AA Numa Scheduler

- Task struct modifications:
  - `tsk->nid` indicates which node on.
  - `tsk->get_child_timeslice` ?

# AA - Data Structures



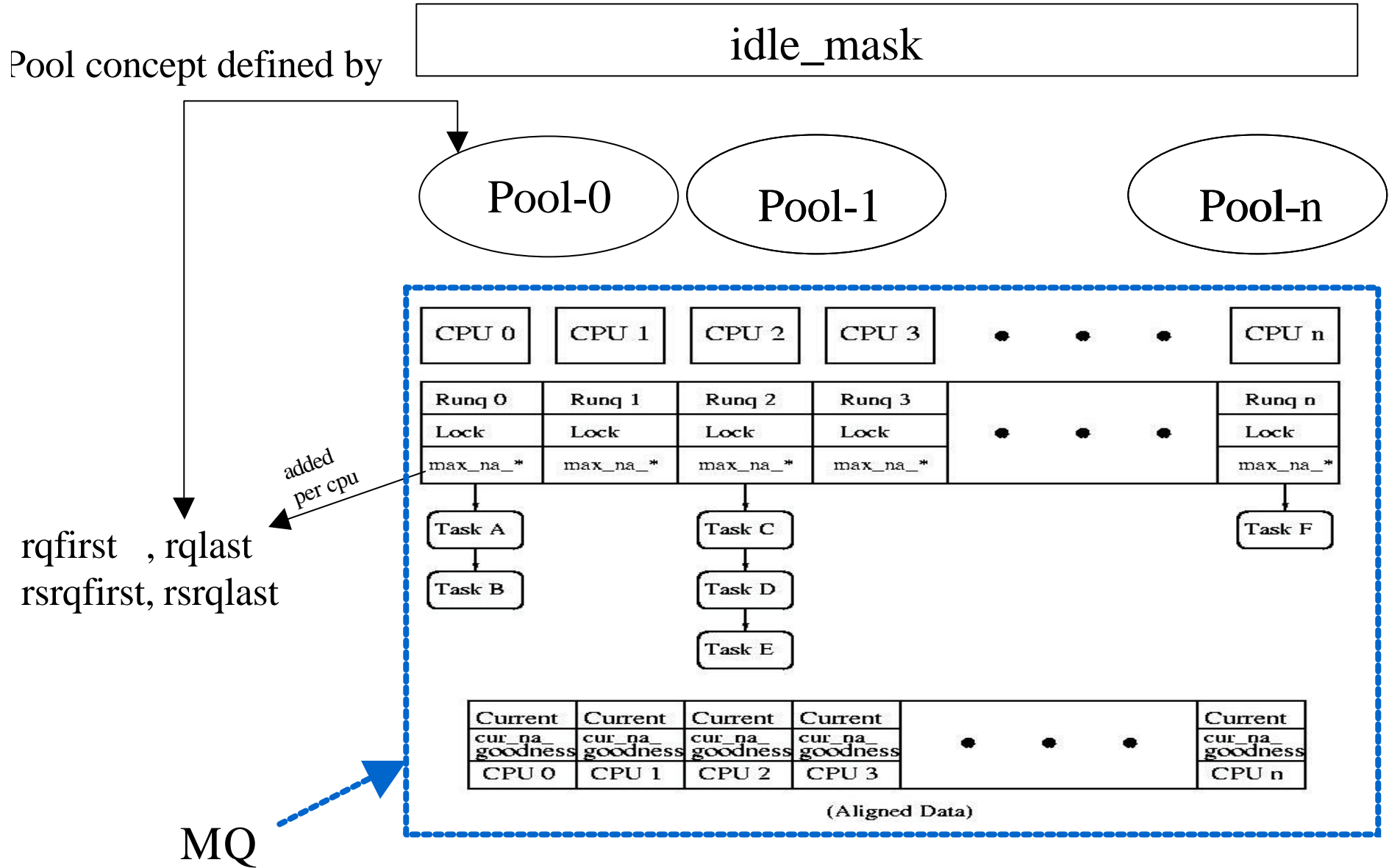
# MQ Overview



Current	Current	Current	Current	•	•	•	Current
cur_na_goodness	cur_na_goodness	cur_na_goodness	cur_na_goodness				cur_na_goodness
CPU 0	CPU 1	CPU 2	CPU 3				CPU n

(Aligned Data)

# PMQS Overview



# reschedule\_idle()

"AA:

- " for all cpus check whether p is on that node and whether we can schedule on that cpu, then determine target (either idle or preemptable task).
- If no target task determined, check all remote node cpus for idle.

"PMQS:

- " check idlemask for local idle
- " check idlemask for remote idle
- " for all cpus in pool check for preemptability and move task to target runqueue

# schedule()

AA :

first search node runqueue

if (c < 0) search remote node queues

if (next->nid != this\_node) move\_task to localqueue

-> remote pull model

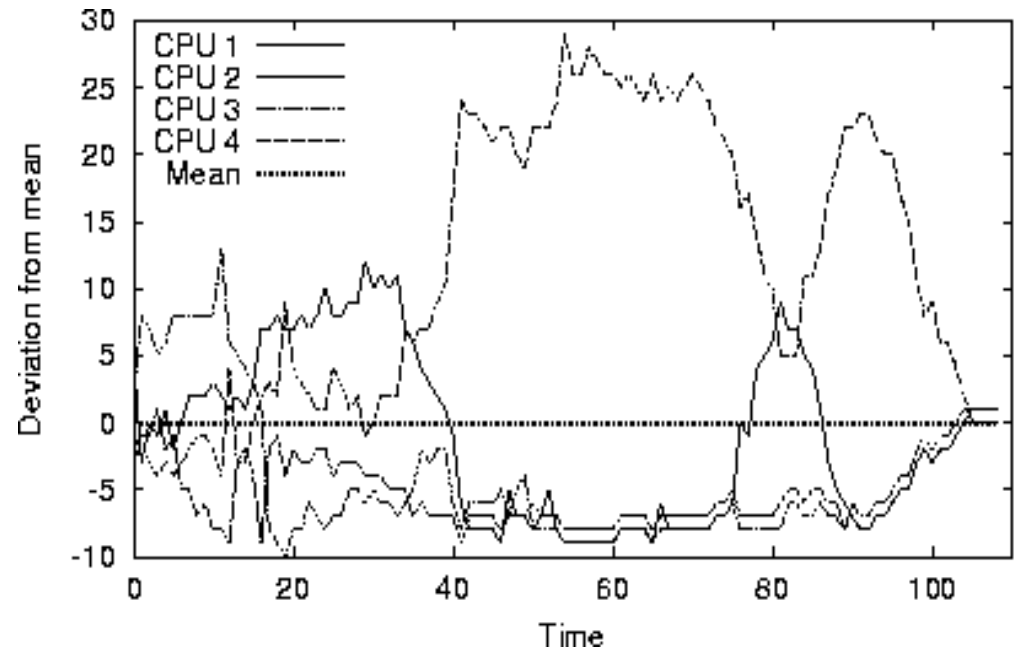
PM QS :

only search in runqueues of localpool, never outside

-> remote push model

# PMQS + LoadBalancing

- Example of RQ length for kernel compile with poolsize=1



- periodically move tasks between pools  
order RQs within pool, order pools  
move from highest RQ in highest pool to  
lowest RQ in lowest pool.

# Comparision ([frival@zk3.dec.com](mailto:frival@zk3.dec.com))

- Ran on 8 cpu Alpha Wildfire (2x4)
  - 2 \* "make -j16 boot" + 8 \* "bonnie++"
  - loadavg: 100 – 200
  - Timed the builds
    - PMQS: loadbalance on 500msec: 1:00 – 1:10 minutes  
loadbalance off: 1:10 - 1:20
    - AANS: 1:50 – 2:10